WHI Hormone Therapy Trial: Estrogen + Progesterone
Baseline and Follow-Up Data Preparation

Introduction

This release of the WHI Hormone Therapy Estrogen + Progesterone Trial baseline and follow-up data includes data collected on study forms, outcomes, results from blood analyses, and computed variables that have been commonly used in data analyses.

Data File Setup

Each data set is provided as a separate fixed length space-delimited ASCII file. The code needed to create a SAS data set from the ASCII file is also provided. Baseline data sets can be found in the *data\epbase\ascii* directory. Follow-up data sets can be found in the *data\epfu\ascii* directory. Each data set is zipped up into a .ZIP file that includes the .DAT raw data file, and the .SAS code file to create the SAS data set. To read the ASCII file into any other statistical program, refer to the INFILE statement in the SAS code file for the order of the variables and to the PROC FORMAT section for the values of all categorical variables.

To use the data you will first need to unzip the .ZIP file for the version of the data set you wish to use.

All data files do not have the same number of records since not every form was completed by each participant. When multiple baseline forms were submitted for a participant, we have included the form with the latest date. The first variable in each file, called ID, is the unique participant identifier that replaces the WHI participant ID. All files are linked by this identifier which MUST be used to merge the data files. The order of the variables after ID matches the order of the questions on the most recent version of the form. In general, computed variables have been added at the end of the appropriate form. The form questions used in the computation of the computed variables have been noted in the variable descriptions. For confidentiality reasons, individual clinical centers are not identifiable.

Each variable has a unique name ranging from two to fifteen characters long. In general, the following extensions were used:

AG	= age
DAYS or DY	= days
EVR	= ever
LST	= last
NUM	= number
NW	= now
OTH	= other
REL	= relative
Y	= year

Dates

No actual dates are included in the data files. All dates have been converted to the number of days since randomization. When only the month and year were recorded, the first day of the month was used to convert the date. A negative number of days indicates the date occurred before randomization. Likewise, a positive number indicates occurrence after randomization.

A small number of baseline forms for required tasks have encounter dates after the date of randomization. We assume these dates reflect edits to the data after the actual randomization occurred.

Data Edits

The built-in features of the data entry system prevented entry of most invalid or impossible data values for all categorical variables. Broad range checks applied to continuous variables have set out-of-range responses to missing. For some variables, like height, weight, and waist circumference the values outside the 1st and 99th percentiles are truncated to the 1st and 99th percentiles. Where this truncation occurs, it is noted in the variable documentation. There still may be values that appear extreme; **it is up to the user to examine all data before proceeding with data analysis.**

Consistency checks between data items on different forms were not done. Therefore, discrepancies do exist. For example, history of breast cancer was collected on both Form 2 and Form 30 and the two data items do not agree exactly. Again it is up to the user to carefully examine the data and determine which values are most appropriate for the specific analyses.

Form Versions

The versions of the data collection forms have changed over time and questions on the forms have been added, deleted, re-ordered and/or modified. To prepare the data for analysis, all questions on each form version were compared to determine if they could be combined into one variable for analysis. In some cases, versions have not been included in the final variables because of incompatibility or because a question was not asked on an early version of a form. This is noted in the data dictionary under usage notes. The text of the question in the data dictionary refers to the latest version of the form. The latest version is assumed to be the final version at the time of this data release.

Missing Data

Missing data can result from a form not being required, a required form not being completed, a particular question on a form not being answered or not required because it was part of a skip pattern, or a question not being asked on all versions of a form. If an entire form is missing for a participant, that participant does NOT have a record in the data file. Missing values in the data files are represented by a single period ("."). The data dictionary gives the number with missing values for all categorical variables. The frequency of missing values could be due to any of the reasons listed above. These frequencies should be confirmed before using the data.

Skip-Patterns

In general, the same skip pattern coding rule has been applied to all data items. If a sub-question is answered inappropriately based on the main question response, it is set to missing. For example, if a sub-question should be answered only if the main question is answered YES, but the main question is answered "No" or "Don't know" or "missing", the sub-question has been set to "missing". If a question is a sub-question, it has been noted as such in the data dictionary. Referring back to the current form should also clarify the question flow. A few exceptions have been made when a large percentage of participants answered the sub-question even though their response to the main question indicates they should have skipped the main sub-question. In these instances, the data in the sub-question was left as is. These exceptions are noted in the usage notes.

Mark-All-That-Apply

Questions involving "mark all that apply" responses have been recoded. Each possible response has been turned into a yes/no variable with a "yes" coded if the response was marked and "no" otherwise. If all possible responses for the question were missing, all possible responses are set to missing. For example, question 16 on Form 20 (medical insurance information) has seven possible responses (codes 1-6 and 8). Seven "yes/no" variables have been created for each participant. If a participant marked 3=Medicare and 8=Other, the variables for the "Medicare" category and "Other" category are coded as "yes", and the variables for the remaining categories are coded as "no".

Current Supplements (Form 45) Data

Data from Form 45 include daily nutrient intake from multivitamins and single supplements and types of supplements taken. The supplement data has been split into two data files as follows: a) nutrients from all supplements, b) types of supplements.

The average intake per day from combination and/or single supplements for 25 nutrients has been calculated. The units of measure for these nutrients match those of the dietary nutrients calculated from the FFQ so that the variables can be summed to yield current nutrient intake from diet and supplements. In calculating these nutrients, the sum has been taken across all types of supplements which can result in extraneous values. After examining the distribution of the nutrient, it may be necessary to truncate extreme values before analysis.

The second file consists of a set of yes/no variables that provide information on the types of supplements taken. For each of the 25 nutrients, a variable was created that indicates if the participant was taking a single supplement containing that nutrient. In addition, variables indicating use of any type of supplement, multivitamins with or without minerals, stress tabs or other combination supplements are included.

FFQ (Form 60) Data

Data from Form 60 include over 100 nutrients that are calculated from participant responses to the FFQ. These nutrient measures are estimates of average daily intake from foods and beverages. Nutrient intake from vitamin and mineral supplements are not included in these totals. Although we provide all nutrients available from the University of Minnesota Nutrition Coding Center nutrient database, there are substantial differences in the reliability of these measures as estimated from an FFQ, where some measures are considered fairly reliable (e.g., percent energy from fat) and others are clearly unreliable (e.g., selenium). For additional information on the WHI FFQ, see: Patterson RE, Kristal AR, Carter RA, Fels-Tinker L, Bolton MP, Agurs-Collins T. Measurement characteristics of the Women's Health Initiative food frequency questionnaire. Annals Epidemiol 1999;9:178-97.

The raw FFQ data (e.g., adjustment question responses, frequencies of consumption, and portion sizes) are not included in this data set.

The nutrient data has been split into four data files, grouped as follows: a) energy, macronutrients, cholesterol, caffeine, fiber, fruits, vegetables, glycemic load; b) vitamins, minerals and carotenoids; c) individual starches, sugars and amino acids, oxalic and phytic acid, and ash; d) individual fatty acids and isoflavones. Consider excluding all nutrient measures for participants with total energy (kcal) less than 600 or greater than 5000 as these energy intake estimates suggest that participants did not complete the FFQ in a reasonable manner.

There are a number of vitamin A related variables in the WHI nutrient dataset that use different units. Investigators using the dataset are advised to refer to the usage notes included in the variable description report to decide which vitamin A variable(s) to use in manuscript analyses.

Blood Results: CBC

The CBC data file includes the results from serum collected at a baseline visit and analyzed at each CC’s local laboratory. All clinical trial and observational study participants were to have serum collected. Data is missing if the lab was unable to process the sample. Values were reported for the following tests: white blood cell count (Kcell/ml), platelet count (Kcell/ml), hematocrit (%) and hemoglobin (gm/dl).

Broad range checks have been applied to the CBC results to exclude biologically implausible values. Extreme values and inconsistencies between results (i.e. hemoglobin and hematocrit) may still exist. **Careful inspection of the data is recommended before using these results in analyses.**

Bone Densitometry Results: BMD

The BMD data files include results from the DXA scans performed at the Clinical Centers participating in the WHI Osteoporosis substudy. Participants with valid results from a hip, spine or whole body scan are included in the data file. These data have been analyzed and monitored by the UCSF DXA Quality Assurance Center before being transferred to the CCC.

In the most recent UCSF DXA QA Report (November 2005), several recommendations were made regarding the data to be used for analysis. They recommended longitudinal and scanner upgrade corrections and provided the necessary correction factors for the following values:

Total hip BMD
Total spine BMD
Whole body BMD
Whole body BMC
Whole body total mass
Whole body total fat
Whole body total percent fat
Whole body total lean
Whole body total fat free mass
Whole body total area

Only the corrected values have been included in the BMD data files.

It was also recommended that "all statistical models with BMD as a dependent variable include scanner (identified by Scanner ID) as a covariate to account for the slight calibration differences between scanners." Variables for the Scanner Ids have been included in the data file, and can be identified by the SAS variable names HIPSID, SPNSID, and WHLSID.

In certain situations, the change in BMD or other DEXA variables between two time points is invalid. Do not compute change if:

1. The two scans were done on different machines, except for calibrated scanner upgrades. Changes are okay between Scanner 3 and Scanner 4, and between Scanner 2 and Scanner 6.

2. The two hip scans were done on different sides of the hip (HIPSDSCN).